中国高通量计算机的自主研发之路

范东睿　叶笑春　包云岗　孙凝晖*
中国科学院计算技术研究所　北京　100190

摘要　随着互联网技术的迅猛发展，高性能计算的主要应用从传统的科学与工程计算为主逐步演变为以数据处理为核心，这给传统高性能计算机体系结构带来巨大挑战的同时，也使高通量计算应运而生。文章从应用特征出发阐述了高通量计算与传统高性能计算的差别，并探讨了高通量计算的基础理论、关键技术，以及中国科学院在高通量计算核心芯片及系统领域的研究成果；以期通过高通量计算机关键技术的研究与突破，为缓解我国核心芯片“卡脖子”的问题，以及为构建智能万物互联时代的新型高性能计算平台作出贡献。

关键词　高性能计算，高通量计算，数据中心，系统熵

DOI　10.16418/j.issn.1000-3045.2019.06.006

近年来，随着互联网技术的迅猛发展，互联网每天产生的数据量呈爆炸式增长。以几个典型公司为例：淘宝网每天交易达数千万笔，其单日数据产生量超过 50 TB；百度每天大约要处理 200 亿次搜索请求，处理数据量达数百 PB；腾讯网日覆盖人数超过 1.5 亿，腾讯视频月总播放量达 800 亿次；Facebook 注册用户超过 20 亿，每月上传的照片达数百亿张。根据国际数据公司（IDC）预测，到 2025 年，全球需要管理的数据量将超过 160 ZB。如何有效对这些数据进行加工将成为一大难题。

在这种背景下，高性能计算的主流应用也从传统的以科学与工程计算为主，逐步演变成以数据处理为核心。然而，由于网络应用及软件技术的不同，需要处理的数据格式和产生速度也各不相同。更甚的是，诸如微博、团购、“秒杀”等网络应用的出现，给大规模数据的实时处理及 QoS（服务质量）提出了更高的要求。因此，互联网技术的普及应用带来的种种新特性给当前的高性能处理器芯片和计算机系统带来了巨大的挑战。

我们都知道，芯片和系统是信息产业发展和安全的根基，尽管我国的信息服务行业发展繁荣，但支撑我国信息行业的核心设施却严重受制于人，特别是关键芯片和核心系统等方面依然面临“卡脖子”的相关问题。当前国内数据中心的中央处理器（CPU）芯片

*通讯作者
资助项目：国家自然科学基金（61732018、61872335）
修改稿收到日期：2019年6月8日

市场几乎被美国的 Intel 和 AMD 两家公司全部瓜分，而加速器芯片则主要由美国的 NVIDIA 公司垄断。核心技术的缺失，使得我国整个信息产业面临着严重的产业安全问题，尤其是当前中美经贸摩擦愈演愈烈，唯有科技自主方可不受制于人。

中国科学院计算技术研究所早在 10 年前就前瞻性地启动了高通量计算机的研究工作。经过多年的科研积累，目前已经在核心芯片、计算机系统等方面形成了诸多创新成果，并已开始逐步投入产业应用。

1 什么是高通量计算机

高性能计算在传统的科学与工程计算类应用中的特点包括：任务单一，负载变化不频繁，单个任务计算量大，以及计算局部性好。而高通量计算在数据中心的应用则主要面向互联网、物联网等新兴场景，其特点是：任务多样，单个任务往往具有流式计算特征；计算量相对不大，但任务的并发数量及数据规模巨大；以及处理要求具有实时性。

传统高性能计算机的研制目标是提高速度，即缩短单个并行计算任务的运行时间；而数据中心类应用系统的目标是高通量，即提高单位时间内任务或数据处理的吞吐量。这种以“算得多”为性能指标的高性能计算机被称为高通量计算机。如果给高通量计算机一个定义，那么可以这么描述：高通量计算机是适用于互联网大数据等新兴应用负载特征的、在强时间约束下能够全局可控地处理高并发请求的新型高性能计算机。其核心特点是对并发性、实时性和确定性的保障。

高通量计算机和传统的高性能计算机在目标应用、计算特征和设计目标等方面都存在明确的区别（表 1）。然而，由于高性能计算由来已久，目前主流的通用计算机和高端计算系统的发展都深受其影响，这也使得当前数据中心主流的计算系统在针对网络服务这种高并发、强实时的高通量应用时表现出诸多不足。为了进一步理解高通量应用对计算机体系结构的需求，我们基于当前主流的高性能服务器（采用 Intel Xeon CPU）对典型高通量应用进行了测试，并且发现了以下一些问题。

（1）缓存资源浪费。CPU 上的共享缓存（cache）缺失率很高，这说明高通量应用与传统高性能计算应用的数据访问特征有明显区别，传统的多级缓存设计并不适合。从面积和功耗的角度来衡量的话，共享缓存作用不大，但却占用了大量的片上面积（在 Intel 的主流服务器芯片中，片上存储所占面积通常高达 30% 以上），产生了大量的功耗。

（2）内存带宽利用率低。CPU 在 70% 以上使用率时的压力测试下，内存带宽的有效使用率通常也不到 10%。这说明，在高通量应用负载下，传统计算机体系结构设计下的内存带宽并没有得到有效利用。

（3）服务质量难以保障。当增加任务的并发负载，使得 CPU 利用率维持在较高水平时，我们发现应用的完成时间迅速拉长，也即系统的尾延迟明显增大，从而导致延迟敏感应用大量失效。因此，在传统服务器系统上，要想获得好的用户体验，必须把硬件利用率维持在较低水平。

通过上述实验结果我们可以看到，现有的高性能计算机系统的设计并不能很好地满足高通量应用的新特性。因此，需要开展新型的高通量计算体系结构的

表 1　高通量计算机与传统高性能计算机对比

	目标应用	计算特征	设计目标
传统高性能计算机	科学与工程计算应用	任务单一，负载变化不频繁，计算量大，计算局部性好	高速度（算得快）
高通量计算机	互联网数据中心应用	任务多样，流式计算特征，数据量大，实时性要求高	高通量（算得多）

研究。

2 高通量计算基础理论

与传统高性能计算以高速度为设计目标相比，高通量计算的核心是追求高通量，即算得多。具体包括3个核心要素，即高吞吐、高利用率、低延迟。

（1）高吞吐。是指单位时间完成的任务数或者响应的请求数要多。对于互联网应用场景来说，数据中心的一个核心挑战是要实时响应海量的并发用户请求。以2018年天猫“双11”全球狂欢节为例，其实时数据处理峰值超过6亿条/秒，支付成功峰值超过30万笔/秒，数据中心必须充分挖掘各种并行性以应对如此巨大的实时并发处理需求。

（2）高利用率。是指计算机系统中的核心部件（如CPU、存储器、网络等）的利用率要高。当前大型数据中心通常包括数十万台甚至百万台服务器，建设资金则高达数十亿甚至百亿美元。然而，为了确保用户的服务质量，现有数据中心不得不将利用率控制在较低水平，因此整体利用率情况很不理想。公开数据显示，2013年谷歌数据中心的平均CPU利用率只有30%[1]，而其他互联网公司运营的数据中心的利用率甚至比该值还要低。可见在现有的架构下，要做到既能实时满足用户处理需求，同时又能达到高的利用率，是非常困难的。

（3）低延迟。指用户请求的响应时间要短。互联网上的大部分在线服务具有明显的实时交互特征，数据中心必须确保在给定的实时性约束条件满足的情况下返回结果，否则会导致服务的失效。比如一些图像识别或者语音翻译之类的人工智能（AI）应用场景，通常要求响应时间在毫秒级别，这对于当前的计算机系统来讲是一个巨大挑战。

针对上述高吞吐、高利用率、低延迟的设计需求，我们提出一个基于“系统熵”的通量分析模型[2]。系统熵主要受延迟的不确定性（波动情况）、资源利用率和吞吐量3个因素影响。简单来讲，系统熵与延迟的波动幅度成正比，与资源利用率以及系统吞吐量成反比。因此，延迟波动越大，系统熵越大；资源利用率越高、吞吐量越大，则系统熵越小。类似于“热力学熵”的用法，我们通过“系统熵”可以反映计算机系统中的易扰动程度或者不确定性。

“熵者，伤也。”高熵系统往往开销大、成本高。相比于高熵系统，低熵系统具有更优的可预测性，能达到更高的效率、更低的成本，也更受用户青睐。曾有人问美国能源部副部长斯蒂文·库宁（Steven Koonin），为什么电能如此受到人们的喜爱？他回答道，因为电力是一种低熵能源。前文提到，为了确保用户服务质量，现有的数据中心的CPU平均利用率很低，一旦利用率提高，其负载性能的波动幅度将迅速增大。因此，当前数据中心计算系统仍然是高熵系统。而高通量计算机的核心目标就是要降低系统熵，也即降低系统的不确定性；以及通过高通量计算机实现提高系统利用率和任务吞吐量的同时，避免应用的性能波动。

3 高通量计算关键技术

针对高通量计算高吞吐、高利用率、低延迟的需求，我们需要把当前计算机体系结构的设计从“速度导向”转向“通量导向”，从而确保计算机系统在满足高吞吐、低延迟的同时还能达到高利用率。针对上述目标，中国科学院计算技术研究所在高通量计算机研制过程中提出了一系列关键技术，包括高通量众核体系结构、高通量片上数据通路、标签化体系结构等。

3.1 高通量众核体系结构

针对高通量应用中的海量并发处理需求，我们提出了Godson-T众核处理器体系结构[3]，以实现任务的高吞吐。相比于传统多核处理器，Godson-T采用众核架构提供丰富的并发处理能力，并在片上网络、片上

存储、同步模型和通信机制等方面采用创新性的设计方法，以实现任务的高吞吐和低延迟。

（1）易扩展片上网络。Godson-T 采用易扩展的二维网格片上网络，同时支持拥塞感知和能耗感知的动态路由算法以实现高并发场景下的片上网络负载均衡，进而确保网络通信的低延迟。

（2）细粒度可配置片上存储。Godson-T 的片上存储支持细粒度可配置，从而更好地适配高通量场景下复杂的数据访问模式，降低延迟。

（3）快速同步机制。我们设计了片上同步管理结构，支持基于数据流的核间细粒度快速同步，相比传统的基于内存的同步机制，性能可获得数量级的提升。

（4）可编程数据通信机制。Godson-T 提出了可编程数据传输引擎结构，可以快速实现数据的水平（片上处理器核之间）和垂直（从内存到片上存储）搬运，实现了数据通信的低延迟。

Godson-T 众核处理器结构受到国际同行的广泛关注，2011 年，处理器领域的知名期刊《微处理器报告》（*Microprocessor Report*）对 Godson-T 的研究成果进行了专门文章报道，并将其选入 2011 年全球十大服务器处理器之一。

3.2 高通量片上数据通路

“通量导向”的处理器数据通路设计也是确保“高吞吐、低延迟”的关键，我们借鉴城市交通管理的思路开展设计。高通量计算在结构特征、资源管理、调度策略等方面都非常类似于城市交通管理，两者的核心特征都是高通量，即在单位时间内完成尽可能多的处理请求，并保证 QoS，表 2 给出了两者的类比情况。

针对应用的新特点，高通量数据通路重点在最基本的数据读取、数据传输（访存通路）和数据处理3个环节进行了创新[4]。

（1）数据读取环节。针对应用中的大量细粒度访存需求，设计了基于硬件的访存请求收集表，通过对大量细粒度访存的收集并批量处理，同时通过时间敏感的收集窗口控制机制，避免长延迟导致的任务失效。

（2）数据传输环节。针对大量细粒度访存的需求，提出了高密度路网的设计，从而提高片上网络的利用率和吞吐量。支持动态通路调整，能根据数据传输的压力，动态调整传输通路配置，提高通路利用率。此外，通过直连快速网络保障关键数据通路的低延迟。

（3）数据处理环节。提出了硬件支持的全局实时任务调度机制，将任务按照优先级及剩余裕度时间进行调度，有效保障任务的 QoS；同时避免对时间裕度不足的失效任务进行调度，从而确保硬件资源的合理利用。

3.3 标签化体系结构

为了在高吞吐、低延迟的同时还能实现高利用率，我们提出了标签化冯·诺依曼体系结构（Labeled von Neumann Architecture，LvNA；图 1）[5-7]。LvNA 的主要思想，是在经典冯·诺依曼体系结构之上增加一套基于标签机制的可编程接口，使得总线与共享硬件部件支持“DIP”能力，即 D—区分（Distinguishing）、

表 2　高通量数据通路与城市交通结构类比

共性特征		结构对比				
高通量数据通路	规模庞大资源繁杂	访存数据收集重组	全局实时调度	快速直连网络	动态路由调度	低宽度高密度网络
城市交通	实时调度需求流量压力大	定制公交	交通指挥控制中心	立交桥	潮汐车道	高密度路网

I—隔离（Isolation）、P—优先化（Prioritizing），从而降低计算机系统内部因资源竞争造成的干扰。

（1）D属性标签机制。在LvNA中，标签将依附于所有的数据访问请求中，用于标识该请求来源于哪一个应用（或应用类别），并随着数据访问请求一同在整个计算机系统中传播。这样，总线和共享硬件部件就可以通过检查数据访问请求的标签来对不同应用（或应用类别）的请求进行区分，从而支持区分属性（D属性）。

（2）I属性标签机制。总线和共享硬件部件可以在对数据访问请求进行来源区分的基础上，对请求所访问的空间资源（如缓存、内存地址空间等）进行隔离，减缓或消除因为空间资源的共享冲突带来的干扰，从而支持隔离属性（I属性）。

（3）P属性标签机制。总线和共享硬件部件可以在对数据访问请求进行来源区分的基础上，对请求所使用的性能资源（如队列、带宽等）进行优先化，减缓或消除因为性能资源的共享冲突带来的干扰，从而支持优先化属性（P属性）。

基于上述标签机制，控制逻辑按照预先设定的规则，以标签为依据对相应的数据访问请求实施不同的性能调控策略。这些性能调控策略是软件可编程的，并且可以做到比传统操作系统的性能调控更为细粒度，从而对延迟敏感型应用会有更优的性能调控效果。

LvNA对硬件的增强并不改动现有指令的语义，因此对软件系统没有侵入性，可以做到无须修改操作系统和应用程序。此外，LvNA不依赖于处理器流水线结构的改动，因而可以适用于任意处理器。

4 高通量计算核心芯片、系统及应用

为了验证高通量计算机在核心芯片和系统等方面的核心技术，中国科学院计算技术研究所先后研制了高通量众核处理器——DPU-m、标签化体系结构——“火苗”，以及高通量计算机系统——“金刚”等，并开展实际应用。

4.1 DPU-m高通量众核处理器

我们完成了DPU-m高通量众核处理器芯片（图2）的设计和流片，芯片基于TSMC 40 nm工艺，主要面向互联网高通量视频处理需求。与数据处理领域的主流芯片Intel的相同工艺芯片相比，能效提升达20余倍。

目前，基于自主技术构建的高通量处理系统在国内外均已开展部署。在国内已经进入国家计算机网络与信息安全管理中心、中国移动、中国联通等重要高通量网络数据监管与分析领域，有效保障了国家信息安全。在国外也已经累计部署数千节点，服务于国家“一带一路”倡议。

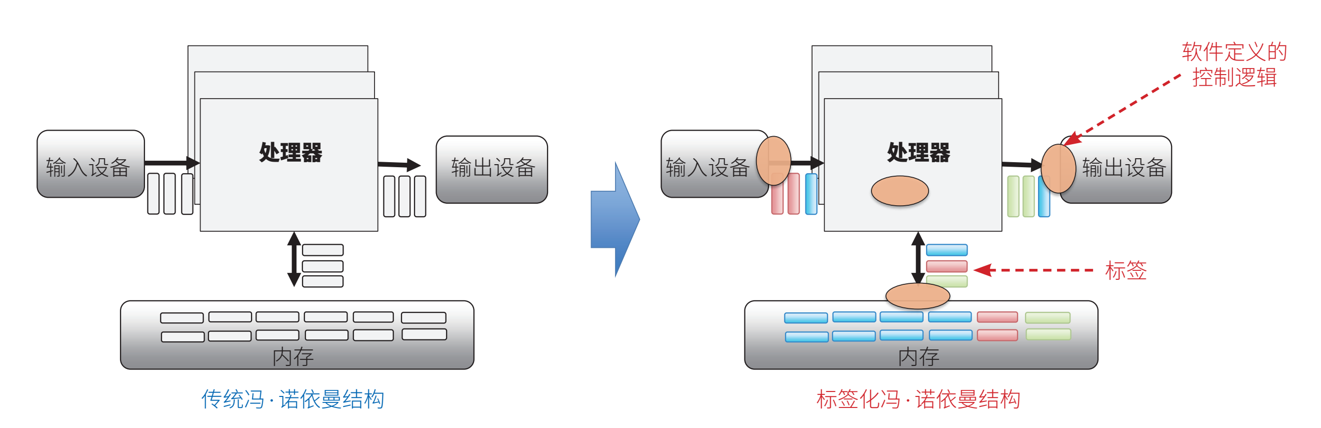


图1　标签化冯·诺依曼体系结构

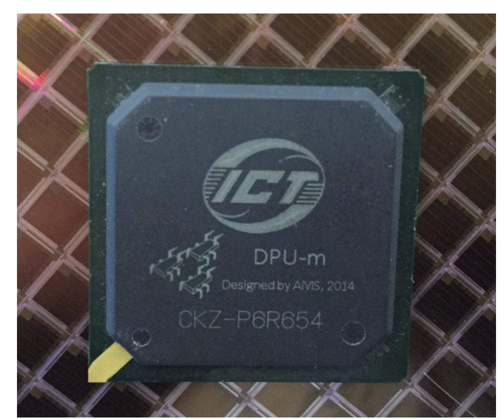




图 2　DPU-m 原型芯片及加速卡

4.2 “火苗”标签化体系结构原型系统

“火苗”原型系统（图 3）是依据 LvNA 实现的 FPGA 原型系统，包括 8 个节点；基于 SiFive 公司 freechips 项目的开源 SoC 实现 Rocketchip，并在其基础上加入了标签化的基础设施以及应用标签的控制平面。该系统已整体达到国际先进水平（美国加州大学伯克利分校于 2018 年 6 月发布同类平台），标签化功能处于国际领先水平。目前，“火苗”原型系统已对外开放，被中国科学院深圳先进技术研究院、清华大学、北京大学、天津大学、大连理工大学、华为海思公司、美国 Clemson 大学等用于前沿研究与产品研发。

4.3 “金刚”高通量计算机

2018 年 10 月，中国科学院计算技术研究所联合北京中科睿芯科技有限公司在中国计算机大会（CNCC）上发布了首台高通量计算机系统——“金刚”（图 4）[8]，该系统集成了该所相关团队在高通量处理器、高通量系统、高通量软件及应用等领域的一系列创新技术，以高吞吐、高利用率、低延迟的特性满足数据中心基础设施建设的新需求，在高并发音视频处理、深度学习等典型应用场景相比传统服务器获得数量级的能效提升。目前，随着高通量计算机系统的成功研制，高通量计算技术将逐步应用到国民经济主战场，贡献于国计民生。

4.4 高通量计算中心建设

当前，城市公共计算基础设施仍以超算中心和云计算中心为主。超算中心采用的是传统高性能计算架构，其核心是“算得快”；以交通工具做类比的话，对应的是飞机，其特点就是速度快、完成时间短。而云计算中心的核心是面对多样化的计算需求实现“算得省”，对应交通工具中的汽车，汽车可以在绝大部分出行场景中都达到成本低和利用率高的目的。然而，飞机和汽车都存在一个明显的局限性：虽然，在流量较低的情况下，两者都能确保较好的服务质量；但是，一旦交通负载快速上升时，就容易造成拥塞，导致完成时间急剧增长，难以保障服务质量。而高通量计算的核心就是要突破上述局限性，在高负载的情况下实现“算得多”，类似于高铁。高铁是目前交通工具中，在高负载、高利用率前提下依然能有效保障用户服务质量的最佳方案。

随着用户出行需求的多样化，交通运输体系也在不断发展完善。类似地，随着应用需求的不断变化，未来城市公共计算基础设施也需要不断发展和完善。面对未来千亿级别端设备带来的新需求，需要提供更高通量、更高智能、更高确定性、更低延迟和更低功耗的计算与传输能力，而高通量计算中心无疑将扮演着越来越重要的角色。

中国科学院计算技术研究所正在开展高通量计算



图3 “火苗”原型系统



图4 “金刚”高通量计算机

中心的建设，按照规划，第一步将先建设1—2个高通量计算的示范中心，然后在全国重点城市开展高通量计算中心建设，逐步实现高通量计算技术与新兴产业的无缝融合。目前，第一个城市高通量计算中心已经选址江苏省盐城市并已开展建设，由中国科学院计算技术研究所团队负责高通量计算中心的整体方案设计、核心设备研制和日常运营。盐城高通量计算中心将重点支持高通量视频处理和人工智能加速，作为服务盐城智能产业升级的核心公共研发平台。

5 总结及建议

经过长期的努力突破，我国在高性能计算机研制方面已经取得一系列令人瞩目的成果。然而我们也看到，不管是传统超算中心，还是新兴互联网数据中心，核心芯片受制于人的现象仍然非常严重。

为了确保我国信息产业的安全可持续发展，有必要以高通量计算等新兴应用场景作为突破口，加强核心芯片和计算系统的自主研发和产业应用，逐步打造自主可控的产业生态。为此，本文提出以下建议。

（1）政策方面，政府明确以高通量计算等为代表的新兴技术的战略定位。一方面，加强以芯片和系统为代表的核心技术专项设置和科研投入；另一方面，加大国家相关部门在高通量计算相关信息基础设施工程的布局和建设，针对国产化自主核心技术在全国挑选重点城市开展试点和验证。

（2）产业方面，整合高通量计算相关优势科研单位、高校及企业，推进相关产业联盟的构建。推动以高通量视频处理、人工智能等为代表的行业应用优先导入产业生态。此外，针对国家“一带一路”倡议，积极探索核心技术产品的出口应用，扩大国际影响力。

参考文献

1 Barroso L A, Clidaras J, Holzle U. The Datacenter as a Computer: An Introduction to the Design of Warehouse-scale Machines. 2nd ed. San Rafael, CA: Morgan & Claypool Publishers. 2013: 96.

2 Sun N, Bao Y, Fan D. The rise of high-throughput computing. Frontiers of Information Technology & Electronic Engineering, 2018, 19(10): 1245-1250.

3 Fan D, Zhang H, Wang D, et al. Godson-T: An Efficient Many-Core Processor Exploring Thread-Level Parallelism. IEEE Micro, 2012, 32(2): 38-47.

4 Fan D, Li W, Ye X, et al. SmarCo: An Efficient Many-Core Processor for High-Throughput Applications in Datacenters//. IEEE International Symposium on High Performance Computer Architecture (HPCA). Washington DC: IEEE Computer Society, 2018: 596-607.

5 Ma J, Sui X, Sun N, et al. Supporting Differentiated Services in Computers via Programmable Architecture for Resourcing-on-Demand (PARD)// Twentieth International Conference on Architectural Support for Programming Languages & Operating Systems. New York: ACM, 2015: 131-143.

6 Bao Y, Wang S. Labeled von Neumann architecture for software-defined cloud. Journal of Computer Science and

Technology, 2017, 32(2): 219-223.

7 包云岗, 范东睿, 陈明宇, 等. "计算机即网络"理念与高通量计算. 中国计算机学会通讯, 2016, 12(3): 10-16

8 新华网. 快速"吞吐"大数据——前瞻计算机"高通量"时代. [2018-10-26]. http://www.xinhuanet.com/politics/2018-10/26/c_1123620065.htm.

Independent Research and Development of High Throughput Computer in China

FAN Dongrui YE Xiaochun BAO Yungang SUN Ninghui*

(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract With the rapid development of Internet technologies, the main applications of high performance computing (HPC) are changing form scientific and engineering applications to those focusing on data processing. This situation poses grave challenges for traditional high performance computing architecture, thus high throughput computing (HTC) comes into being. This paper explains the difference between HPC and HTC according to the application features, as well as introduces the basic theory and key technologies of HTC. We also show the research results of HTC chips and systems. Through the breakthrough of above HTC key technologies, it is expected to relieve the bottleneck of core chips, and make due contribution to the China's new high performance platform in the era of intelligent Internet of Things (IoT).

Keywords high performance computing (HPC), high throughput computing (HTC), datacenter, sys-entropy



范东睿 中国科学院计算技术研究所高通量计算机研究中心主任、研究员，中国科学院大学岗位教授。曾获中组部"万人计划"领军人才和CCF-IEEE CS青年科学家等荣誉。主要研究领域包括高通量计算、众核处理器体系结构等。E-mail: fandr@ict.ac.cn

FAN Dongrui Ph.D., Professor. He is currently the Director of High Throughput Computer Research Center, Institute of Computing Technology, Chinese Academy of Sciences (CAS), and Professor of University of Chinese Academy of Sciences. Dr. Fan holds several high profile honors, including National High-level Personnel of Special Support Program, CCF-IEEE CS Young Computer Scientist Award. His main research interests include high throughput computing and many-core processor architecture.
E-mail: fandr@ict.ac.cn

*Corresponding author



孙凝晖　中国科学院计算技术研究所所长、研究员，计算机体系结构国家重点实验室主任、学术委员会副主任，中国计算机学会高性能计算专业委员会主任，中国计算机学会副理事长，《计算机学报》主编，中国科学院大学计算机与控制学院副院长，中国科学院信息科技领域发展路线图战略研究专家组组长。主要研究领域包括高性能计算、高通量计算。E-mail: snh@ict.ac.cn

SUN Ninghui　Ph.D., Professor. Currently, he is the Director of Institute of Computing Technology, Chinese Academy of Sciences (CAS). He is also the Director of State Key Laboratory of Computer Architecture, and Vice Chairman of the Academic Committee. Dr. Sun is the Vice President of China Computer Federation (CCF), and the Director of CCF TCHPC. He serves as editor-in-chief of *Chinese Journal of Computers*. He is the Vice Dean of School of Computer Science and Technology, University of Chinese Academy of Sciences. Currently, Dr. Sun is the leader of the expert group on strategic studies of information technology development roadmap in CAS. His main research interests include high performance computing and high throughput computing. E-mail: snh@ict.ac.cn

■责任编辑：文彦杰

参考文献（双语版）

1 Barroso L A, Clidaras J, Hölzle U. The datacenter as a computer: An introduction to the design of warehouse-scale machines, 2nd ed. Synthesis Lectures on Computer Architecture, 2013, 8(3): 1-154.

2 Sun N H, Bao Y G, Fan D R. The rise of high-throughput computing. Frontiers of Information Technology & Electronic Engineering, 2018, 19(10): 1245-1250.

3 Fan D, Zhang H, Wang D, et al. Godson-T: an efficient many-core processor exploring thread-level parallelism. IEEE Micro, 2012, 32(2): 38-47.

4 Fan D, Li W, Ye X, et al. SmarCo: An efficient many-core processor for high-throughput applications in datacenters//. IEEE International Symposium on High Performance Computer Architecture (HPCA). Washington DC: IEEE Computer Society, 2018: 596-607.

5 Ma J, Sui X, Sun N, et al. Supporting differentiated services in computers via programmable architecture for resourcing-on-demand (PARD)// Twentieth International Conference on Architectural Support for Programming Languages & Operating Systems. New York: ACM, 2015: 131-143.

6 Bao Y G, Wang S. Labeled von Neumann architecture for software-defined cloud. Journal of Computer Science and Technology, 2017, 32(2): 219-223.

7 包云岗, 范东睿, 陈明宇, 等. “计算机即网络”理念与高通量计算. 中国计算机学会通讯, 2016, 12(3): 10-16.
Bao Y G, Fan D R, Chen M Y, et al. “Computer is network” concept and high-throughput computing. Communications of China Computer Federation, 2016, 12(3): 10-16. (in Chinese)

8 新华网. 快速“吞吐”大数据——前瞻计算机“高通量”时代. [2018-10-26]. http://www.xinhuanet.com/politics/2018-10/26/c_1123620065.htm.
XINHUANET. Fast throughput big data—Looking forward to the era of “high throughput” of computers. [2018-10-26]. http://www.xinhuanet.com/politics/2018-10/26/c_1123620065.htm. (in Chinese)